Acquiring Verb Classes Through Bottom-Up Semantic Verb Clustering

Olga Majewska, Diana McCarthy, Ivan Vulić and Anna Korhonen
University of Cambridge
Language Technology Lab
Dept. of Theoretical & Applied Linguistics
Cambridge CB3 9DA, United Kingdom
om304@cam.ac.uk, diana@dianamccarthy.co.uk, iv250@cam.ac.uk, alk23@cam.ac.uk

Abstract

In this paper, we present the first analysis of bottom-up manual semantic clustering of verbs in three languages, English, Polish and Croatian. Verb classes including syntactic and semantic information have been shown to support many NLP tasks by allowing abstraction from individual words and thereby alleviating data sparseness. The availability of such classifications is however still non-existent or limited in most languages. While a range of automatic verb classification approaches have been proposed, high-quality resources and gold standards are needed for evaluation and to improve the performance of NLP systems. We investigate whether semantic verb classes in three different languages can be reliably obtained from native speakers without linguistics training. The analysis of inter-annotator agreement shows an encouraging degree of overlap in the classifications produced for each language individually, as well as across all three languages. Comparative examination of the resultant classifications provides interesting insights into cross-linguistic semantic commonalities and patterns of ambiguity.

Keywords: verb classes, semantic clustering, multilingual NLP

1. Introduction

With the recent advances in automatic lexical acquisition, the need for high-quality evaluation resources is ever growing. Due to the pivotal role played by verbs in sentence structure, the problem of creation of verbal classifications has attracted a lot of attention in natural language processing (NLP). Different approaches to creation of verbal classifications have been proposed, varying with regard to the guiding criteria by which the class architecture is organised, prioritising semantic (WordNet (Miller, 1995; Fellbaum, 1998), FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005)) or syntactic information (COMLEX (Grishman et al., 1994), VALEX (Korhonen et al., 2006)), or combining the two (Levin, 1993; Kipper et al., 2000; Kipper Schuler, 2005). Kipper Schuler's (2005) VerbNet, grouping English verbs into classes defined by shared meaning components and syntactic behaviour, is one of the richest lexical verb resources currently available, and its utility in various NLP applications has been repeatedly demonstrated (Rios et al., 2011; Windisch Brown et al., 2011; Schmitz et al., 2012; Lippincott et al., 2013; Bailey et al., 2015).

However, creation of a similar resource from scratch, drawing simultaneously on semantic and syntactic criteria, is a challenging and time-consuming task when attempted by annotators without theoretical linguistics background (Majewska et al., 2017). A number of approaches to automatic verb classification have been proposed (Joanis et al., 2008; Sun et al., 2010; Falk et al., 2012; Kawahara et al., 2014; Scarton et al., 2014; Peterson et al., 2016; Vulić et al., 2017), allowing to minimise the time required and eliminate the need to employ trained lexicographers. However, evaluation of such systems relies on the availability of gold standard classes, and these are still lacking for a great majority of languages.

In light of these challenges and the high demand for verbal resources, this paper investigates whether semantic verb classes can be reliably acquired from non-expert native speakers based solely on verb semantics and following simple instructions, which, to the best of our knowledge, is the first evaluation of this approach. Drawing on the hypothesis that syntactic and semantic behaviour of verbs are tightly interrelated (Pinker, 2013; Jackendoff, 1992; Levin, 1993), we simplify the classification task by eliminating the need to refer to explicit syntactic knowledge and assess whether intuitive native-speaker perception of closeness of verb meaning provides enough guidance to produce consistent verb classifications. This will allow us in future work to examine the relationship between semantics and syntactic behaviour of the class members. Previous classifications have used syntactic behaviour to guide the construction of verb classification but this necessitates linguistic training.

In order to examine the potential of manual semantic clustering in different languages, we carried out verb clustering experiments with native speakers of English, Polish, and Croatian. We describe the set-up of the task in Section 2. Subsequently, we analyse the inter-annotator agreement for each language individually and examine the overlap between classes cross-lingually. Section 3 includes the results of this evaluation. Finally, in Section 4, we discuss observations made with respect to the easily classifiable verbs and those which caused problems in all the languages considered, which shed light on cross-linguistic semantic commonalities and polysemy patterns.

2. The Semantic Verb Clustering Task

The task involved a group of 8 native-speaker participants without formal linguistics training, 3 annotators for English and Polish, and 2 in Croatian, who performed soft clustering of a sample of verbs in their native language based on the verbs' semantic similarity. The verb samples were created as follows: first, a sample of 267 English verbs was au-

	English				Polish				Croatian		
	A1	A2	A3	**Ave**	A1	A2	A3	**Ave**	A1	A2	**Ave**
Number of classes	61	77	58	**65.3**	47	46	35	**42.7**	88	76	**82.0**
Average class size	4.4	3.5	4.6	**4.1**	5.7	5.8	7.6	**6.3**	3.0	3.5	**3.3**
Time spent [hours]	2	1	3	**2.0**	3	3	3	**3.0**	3	2	**2.5**

Table 1: Results and statistics of semantic clustering of 267 verbs for English, Polish, and Croatian, for each annotator (A1-A3) and the average scores for each language (Ave)

tomatically extracted from the pool of SimVerb-3500 (Gerz et al., 2016) verb types. The verbs were sampled so as to ensure that the top 34 VerbNet classes (according to the number of verbs in the class) from SimVerb-3500 are represented by at least 5 member verbs each, to guarantee 'clusterability' of the verbs presented to the annotators. Next, the English sample was translated by native-speaker translators into Polish and Croatian, and the three samples were manually inspected.

Before the start of the task, the annotators were provided with instructions (Appendix) and a list of 267 verbs in a text file, presented in random order, one word in each line. Since the goal was to keep the task as simple as possible for participants without linguistics training, the annotation guidelines were intentionally minimal: they instructed the annotators to put verbs together using a spreadsheet program (e.g. Microsoft Excel) so as to form groups containing verbs that are used to express similar or related meanings. The groups could vary in size, but annotators were asked to aim for at least 3-5 members. A verb could be put in more than one class (e.g. when it had several different meanings), and any verb which did not seem to fit with any group could be placed in a 'Miscellaneous' class. Annotators were encouraged to make a note of any relationship links between groups where they felt the meanings of member verbs were in some way related, e.g. a bidirectional link between similar groups, or a unidirectional link between a broader class and its subclass(es).

3. Results and Inter-Annotator Agreement

The results and statistics of the semantic clustering task for each annotator individually and across annotators, in each of the three languages considered, are reported in Table 1. On average, it took 2.5 hours to complete the task across all annotators, ranging from 1 to 3 hours. The average number of classes obtained was 65.3 for English, 42.7 for Polish, and 82 for Croatian, with class size ranging from the average of 3.3 member verbs in Croatian to 6.3 members in Polish.

3.1. Percentage IAA

In order to measure the overlap between classifications produced by annotators for each language individually and across languages, we calculate percentage inter-annotator agreement (% IAA) for all pairings of verbs. First, we extract all the pairs of verbs on which there is perfect agreement (i.e. all annotators either grouped them together or not), for each of the languages independently, and compute the ratio of observed agreement pairs to all the possible pairings of verbs. Subsequently, we repeat the same procedure for all the English, Polish and Croatian annotators together.

	English	**Polish**	**Croatian**	**All**
% IAA	88.5%	92.5%	97.8%	79.9%

Table 2: The percentage inter-annotator agreement calculated for all possible pairings of verbs, for each language individually and across the three languages

The computations yield a high inter-annotator agreement score for each of the languages, with 88.5% observed for English, 92.5% in Polish, and 97.8% in Croatian (Table 2). The percent inter-annotator agreement calculated across all three languages is 79.9%. It must be noted that the very high agreement score obtained for Croatian, compared to the other two languages, is likely to be due to the smaller average class size. Since many Croatian classes included as few as 3 member verbs, there was a large number of pairs of verbs which were not classified together. Whenever the annotators agreed on not putting two verbs together, that pair constituted an 'agreement' pair for the purposes of inter-annotator agreement calculation. The smaller classes gave rise to the somewhat inflated % IAA score for Croatian because of the larger number of true negatives (verbs that are correctly found not to go in the same class). Its inclusion of true negatives gives % IAA rather high scores generally. In order to address this issue, in the following section (3.2) we calculate inter-annotator agreement using a different evaluation metric, Fuzzy B-Cubed for overlapping clusters (Amigó et al., 2009; Jurgens and Klapaftis, 2013)[1], which avoids the problem of inflation due to scoring true negatives.

3.2. B-Cubed for Overlapping Clusters

In our verb-clustering task, the total number of classes was left unspecified and the annotators were free to put a single verb in as many different classes as they felt was appropriate, whenever they recognised it had more than one distinct sense. In order to adequately evaluate the results, the evaluation measure applied to our data had to be able to accommodate these characteristics of the task. We chose the B-Cubed metric (Bagga and Baldwin, 1998) extended by Amigó et al. (2009) to compare overlapping clusters, and

[1]We used the Fuzzy B-Cubed implementation of Jurgens and Klapaftis (2013) but did not associate the clusters with weights, and therefore the metric is equivalent to that of Amigó et al. (2009).

	Average B-Cubed
English	0.262
Polish	0.338
Croatian	0.172
All	0.205
1c1inst	0.0
All-instances, One class	0.069

Table 3: The average B-Cubed F-score (i.e. harmonic mean of B-Cubed precision and recall) calculated for all possible pairings of annotators, for each language individually and across the three languages, and for two SemEval baselines: 1c1inst and All-instances, One class

by Jurgens and Klapaftis (2013) to fuzzy clusters, used to evaluate the performance of Word Sense Induction systems in SemEval tasks (Jurgens and Klapaftis, 2013).

The B-Cubed metrics (B-Cubed precision and recall) compare two clusterings (say, X and Y) at the item level: for an item *i*, precision measures how many items sharing a cluster with *i* in clustering X are placed in its cluster in clustering Y; whereas B-Cubed recall measures how many items sharing a cluster with *i* in Y are also placed in its cluster in X, with the final B-Cubed score equivalent to the harmonic mean of the two values.

In our task, rather than comparing each clustering against a gold-standard set of classes, we calculate the B-Cubed score for each pair of clusterings produced by the annotators, for each language individually and across all three languages. We report the results of this evaluation in Table 3. The highest agreement score is observed for Polish, where the average B-Cubed F-score is 0.338. Less overlap was found between the clusterings produced for English (0.262), with the lowest B-Cubed F-score obtained for the Croatian clusterings (0.172). The low score reported for Croatian is especially noteworthy in the light of the inflated percent agreement result reported in section 3.1. With percent agreement computed for every possible pairing of verbs, based on a binary choice between two verbs being either clustered together or kept separate, the two annotators seemed to agree in a vast majority of their clustering decisions. Applying an alternative evaluation metric allows us to identify the bias from scoring true negatives, i.e. all cases in which the annotators agreed that two verbs should not be clustered together. As predicted, this inflation is particularly high in the case of Croatian due to the small average class size compared to the other two languages. Indeed, manual inspection of the classes produced by the Croatian annotators shows that in some cases the minimum class size of 3-5 members recommended by the guidelines was not adhered to.

The average B-Cubed F-score calculated for all possible pairings of annotators across the three languages (using translational equivalents for cross-lingual comparisons) is 0.205. Notably, the average cross-lingual agreement score is higher than the value obtained for Croatian itself, which suggests a promising degree of overlap between English and Polish classes (the average B-Cubed F-score for these two languages is 0.237).

Keeping in mind the differences in the nature of the present task and a Word Sense Induction task (which can be seen as an example of unsupervised clustering, with usages of a word grouped into clusters, each representing uses of the same meaning (Jurgens and Klapaftis, 2013)), comparing our results against the scores obtained by the SemEval participating systems may help interpret the reported values. Overall, the top-performing system surpasses our highest result for Polish (scoring 0.483), on the other hand, in the multi-sense setting (i.e. on instances labeled with multiple senses), the best performing system achieves the B-Cubed score of 0.134, a result below the lowest agreement score in our task.

In order to make the comparison more meaningful, we calculate two SemEval baselines for our task: (1) 1c1inst, where each instance is assigned to its own class, and (2) All-instances, One class, which assigns all instances to a single class. The result for the first baseline, 0.0, is the same as in SemEval, and a natural consequence of B-Cubed since there are no pairs within a class. However, while the overall performance of the All-instances, One sense baseline in SemEval surpasses its best participating system (achieving the score of 0.623), the result for this baseline on our verb clustering is much lower (0.069), suggesting the task is significantly more difficult, due to the high number of clusters. And yet, despite the greater difficulty of the task, the agreement between our annotators exceeds the performance of the baselines, which is an encouraging outcome.

As noted earlier, our verb clustering task and the SemEval task are different. The SemEval annotation was performed using predefined senses for graded-tagging (on a Likert scale) and the systems' clusters were compared to clusters induced from these graded sense-annotations. Since the senses for consideration by the annotators were defined in WordNet this is not comparable with our task of clustering verbs. Our task allowed for complete flexibility in the number of classes, which resulted in varying levels of granularity (e.g. Croatian classifications had up to 88 clusters, while the smallest Polish clustering had 35), and a higher number of clusters overall with respect to the SemEval task. The different set-ups of the two tasks entail different levels of difficulty and hence, different agreement scores, and this is reflected in the results obtained for the same baselines discussed above. What is important, our analysis constitutes the first attempt at measuring agreement on clustering of verbs performed by humans.

The encouraging degree of overlap observed between the classifications produced in our manual clustering task, particularly for Polish and English, suggests that there are consistent patterns in how humans group verbs based on their semantic similarity, not only in each language independently, but also across languages from different language families. Collecting more classification data for Croatian, while controlling for class size (as per the minimum class size stated in the guidelines), will allow to verify whether the lower B-Cubed score reported for that language has to do with the peculiarities of the collected data or is indicative of a general greater difficulty in classifying verbs in

Croatian with respect to the other two languages. Extending the experiments to other diverse languages will allow to investigate even further the extent to which those regularities are observed cross-linguistically; however, these are already promising inter-annotator agreement results for a multilingual semantic task.

3.3. Cross-Linguistic Areas of Overlap

Manual inspection of the resultant classes from all annotators allows us to observe what class types and semantic domains are shared by the three languages. Five classes emerge which share the core of at least 2 member verbs across annotators in all three languages (with extra members added by some annotators) and can be described with the following labels (denoting 'verbs of_'): 'looking', 'cooking', 'existing', 'movement in water', 'emitting sound'. The total of 30 classes can be identified where the core of at least 2 member verbs is shared by at least two languages (by all annotators), and whose members belong to the same semantic domains across languages (but with more variation in specific member verbs recorded by individual annotators). In section 4, we look more closely at semantic patterns observable in all three languages and discuss which aspects of verb meaning make the classification task consistently easier or harder, regardless of the language in question.

4. Analysis and Discussion

Despite the encouraging inter-annotator agreement scores, several issues affecting the agreement and overlap between the resultant classifications could be observed. First of all, as the task did not impose a fixed number of classes, the levels of granularity varied between annotators: the difference between the minimum and maximum number of classes equals 12 for Polish and Croatian, and 19 for English. This discrepancy is even more noticeable across languages: while Polish annotators grouped verbs into 35-47 classes, Croatian classifications comprise between 76-88 verb classes. As the task consisted in grouping verbs into flat classes, the resultant classifications do not capture hierarchical relationships between verb groups (these could, however, be signalled as 'relationship links', as noted above). Therefore, potential inclusion of one class by another (e.g. in the case of 'movement' verbs, which in Croatian are split into two small classes, depending on the medium (water vs ground), and are grouped together in one broader 'movement' class in Polish (*swim, dive, walk, crawl*)), is interpreted as class disjunction in automatic pairwise evaluation, which results in a lower overlap between the classifications. What is more, in some cases distinct patterns of ambiguity in the languages considered resulted in different clustering decisions: for example, while in English two senses of the verb 'shine' (i.e. emit light and polish (a shoe)) were considered, resulting in pairings 'shine'-'glow' and 'shine'-'brush', only the former sense is available in Polish and Croatian.

4.1. Problematic and Easily Classifiable Verbs

In order to investigate whether some verbs are inherently easier or harder to classify, and examine to what extent this is observed across languages, we extracted all the pairs of verbs on which there is perfect agreement and those on which the annotators disagreed for each language individually, and examined the overlap between these groups of verbs across the three languages. This allowed to identify 72 'problematic' and 24 'easy' verbs, shared by the three languages. Manual inspection of these groups let us make a number of observations regarding the aspects of verb meaning which pose problems or make them easier for humans to classify, regardless of the language considered.

4.1.1. 'Problematic' Verbs

Most of the verbs which ended up in the 'problematic' group share the characteristic of having a broad, vague or abstract meaning, sometimes with several related senses which allow them to appear in a number of slightly different contexts. For example, annotators in all three languages disagreed on how to classify verbs such as *affect, treat, engage* or *spare*. What is more, some display a degree of semantic vacuity, that is, have little semantic content of their own and tend to express a more precise meaning when combined with some other word (e.g. a noun), with which they form a predicate, such as *make* or *have*, examples of the so-called 'light verbs' (Jespersen, 2013). Inspection of the VerbNet classes from which the 'problematic' verbs were sampled revealed that the 'Change of State' class (45) is particularly often represented. Although verbs such as *slip, vary* and *tumble* belong to the same VerbNet subclass (45.6-1, 'calibratable change of state'), their meanings are not intuitively similar. Moreover, each has several senses, which is reflected in the fact that each participates in a number of distinct VerbNet classes. Understandably, this results in more variation in clustering decisions, as different annotators are likely to take different verb senses into consideration, and consequently, produce divergent classifications.

4.1.2. 'Easy' Verbs

Verbs which lend themselves better to manual semantic classification are those with narrow, concrete meanings, for example, verbs describing sounds (*chirp, buzz, roar*) or those belonging to a clearly defined semantic field, e.g. 'cooking' verbs (*fry, bake, cook*). Synonymous verbs such as *study* and *examine*, or *observe* and *stare*, were also among those on which the same clustering decisions were made across annotators, in all three languages. Interestingly, there was full agreement on antonymous pairs such as *vanish* and *appear*, which were consistently grouped together in all languages. As discussed in lexical semantics literature (Cruse, 1986), antonyms have a paradoxical nature: on the one hand, they constitute the two opposites of a meaning continuum, and therefore could be seen as semantically remote, on the other hand, they are paradigmatically similar, having almost identical distributions, and hence seem closely related. Despite these conflicting properties of antonyms, humans seem to intuitively recognise their relatedness and consistently group them together, as semantically similar. The perception of relatedness overrides the sense of 'oppositeness' and being maximally distant along a dimension of meaning, and opposites end up clustered together. This regularity is observed in the case of pairs of relational antonyms, i.e. verbs which describe an

event from opposite points of view, for example, *lend* and *borrow*, which differ along only one dimension of meaning, that is, the direction of the action (the object of the verb either travels away from the participant (*A lends something to B*) or towards the participant (*B borrows something from A)*), and are essentially identical with regard to all other features, which makes them appear semantically close.

4.2. Semantic Similarity versus Relatedness

The importance of distinguishing between the concepts of semantic similarity (e.g. *cup* and *mug*) and relatedness (e.g. *coffee* and *cup*) has been noted in the literature (Hill et al., 2015), and the analysis of our data provides more evidence illustrating the influence of loose association on how humans conceptualize similarity between words, and the difficulty of keeping similarity and relatedness apart. In all three languages we can observe instances of what can be described as a 'storyline approach' to judging semantic similarity and verb classification. This is particularly noticeable in Croatian classifications, where several classes formed by the annotators group verbs describing quite different actions, linked via loose thematic ties: (1) *marry, conquer, approach, move*, where putting semantically dissimilar verbs *marry* and *move* together seems to suggest an underlying 'storyline' with courtship leading to marriage and moving house; (2) *visit, communicate, treat, operate*, where the association of verbs *visit* with *treat* and *operate* brings to mind a hospital visit, or (3) *finish, frame, announce, submit*, which can be seen as belonging to an 'academic' thematic domain. Relying on association rather than actual consideration of semantic components of verbs' meaning is visible in the cases where a class contains verbs which express a consequence of the action or state described by other verbs in the same class, e.g. *glow, shine, squint* in one of the Polish classifications, with 'squinting' being a reaction to 'glowing' or 'shining', or *ache, hurt, kick, rub, cry* in Croatian. Although verb groupings in which loosely thematically related verbs are classified together are in the minority, their presence in our data suggests that, in order to obtain classes based solely on semantic similarity judgments, unbiased by loose association, the annotation guidelines should explain the similarity-relatedness distinction and instruct the annotators accordingly.

4.3. Polysemy

An in-depth investigation of the resultant classes also offers an insight into the patterns of polysemy in the three languages considered. In our task, the annotators could accommodate a verb's ambiguity by placing it in several different classes, putting each of its distinct senses in a separate cluster. However, since the annotators were provided with just word forms and the senses were not specified a priori, there were some discrepancies in which senses were identified, across annotators and, expectedly, across the different languages, which led to a lower cross-lingual agreement in the resultant classes. For example, the Croatian translation of the English verb 'to vary', *odstupati*, expresses not only the sense of 'differing', but also 'withdrawing', unavailable in English or Polish, which explains why it was placed in the same class with *move* and *renounce* only by the Croatian annotators. Analogously, the Croatian equivalent of *remark* (*primijetiti*), ambiguous between senses 'to comment' and 'to notice', ended up together with verbs such as *look, stare, observe*, while in Polish and English it was grouped with verbs of 'communicating'. Similarly, the Polish translation equivalent of the verb *weave* (*pleść)* is ambiguous between two senses, 'to interlace' and 'to blabber', and was grouped both with *join* and *combine*, and with *tell* and *communicate*, while no such ambiguity was recorded in English and Croatian. Finally, while two senses of the verb *sway* ('to move rhythmically from side to side' and 'to control or influence') are available in English, only the former is recognised in the Polish and Croatian classifications and its translation equivalents are never grouped together with verbs such as *convince* or *persuade*, as it is the case in English.

In a task such as ours, where guidelines were intentionally restricted, so as to avoid imposing any preconceived semantic categories or classification structure onto the annotators and elicit possibly spontaneous similarity judgments, such discrepancies in detecting ambiguity are inevitable. In order to have more control over which sense of a given verb is taken into consideration in the clustering task, word senses rather than word forms would have to be provided at the start of the task. Such a set-up would also allow comparison of the elicited classes with the existing multilingual sense inventories, like Open Multilingual WordNet (Bond and Foster, 2013) or BabelNet (Navigli and Ponzetto, 2012). Since the aim of the present study was to elicit judgments on basic word forms, without any guidance as to the different word senses available, such comparisons are beyond the scope of this study; however, in future work we intend to extend this analysis and compare our findings against the resources available.

5. Conclusion

We have presented the first cross-lingual analysis and evaluation of semantic clustering of verbs by non-expert human annotators. The inter-annotator agreement scores reported for English, Polish, and Croatian are encouraging and demonstrate that verbs can be reliably classified by humans without linguistics background. What is important, this suggests that there is potential to create verb classifications starting from a simple, purely semantic task. Moreover, the degree of overlap in the resultant classifications observed across languages implies that there are cross-linguistic commonalities and shared meaning components governing the semantic organisation of verbs. A cross-lingual scrutiny of low-agreement verbs and those on which annotators made identical clustering decisions, allowed us to investigate to what extent the same verbs are problematic and whether some verbs are inherently easier to classify. Manual inspection of the thus identified 'easy' and 'problematic' verbs provided interesting insights into the aspects which may affect 'clusterability' of verbs across different languages. The present study opens up several avenues for future work. First of all, we would like to extend the study to other languages from different language families, as well as test the applicability of our bottom-up semantic-based

approach to larger verb samples. This would allow us to expand the cross-lingual analysis of the semantic classes obtained and their underlying properties, and investigate how our findings about 'easy' and 'problematic' verbs are reflected in other language resources and corpora. What is more, a comparison of our classifications against VerbNet could provide interesting insights into where speakers' intuitions about word usage most diverge from semantic boundaries drawn by lexicographers, and how the existing verb resources could be improved to better reflect speakers' perceptions about verbs' semantic characteristics and behaviour. Moreover, a comparative analysis of our data against an output of an automatic clustering algorithm would allow us to investigate whether the manual classification task can be (partly) substituted with a semi-automatic one, with an initial rough clustering based on verbs' distributional properties extracted from a large corpus, and subsequently verified by a human annotator.

6. Acknowledgements

We gratefully acknowledge the funding support of the Economic and Social Research Council [PhD Award Number ES/J500033/1] and the European Research Council (ERC) [Consolidator Grant 648909)]. We thank Daniela Gerz for helpful discussions on this work and the three anonymous reviewers for their comments and suggestions.

7. Bibliographical References

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada, Spain.

Bailey, D., Lierler, Y., and Susman, B. (2015). Prepositional phrase attachment problem revisited: how verbnet can help. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of ACL-COLING*, pages 86–90.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of ACL*, pages 1352–1362.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Falk, I., Gardent, C., and Lamirel, J.-C. (2012). Classifying french verbs using french and english lexical resources. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 854–863. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet*.

Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Grishman, R., Macleod, C., and Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. In *Proceedings of COLING*, pages 268–272.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jackendoff, R. S. (1992). *Semantic structures*.

Jespersen, O. (2013). *A Modern English Grammar on Historical Principles: Volume 5, Syntax (fourth Volume)*, volume 5. Routledge.

Joanis, E., Stevenson, S., and James, D. (2008). A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

Jurgens, D. and Klapaftis, I. P. (2013). Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *SemEval@ NAACL-HLT*, pages 290–299.

Kawahara, D., Peterson, D., and Palmer, M. (2014). A step-wise usage-based method for inducing polysemy-aware verb classes. In *ACL (1)*, pages 1030–1040.

Kipper, K., Dang, H. T., and Palmer, M. S. (2000). Class-based construction of a verb lexicon. In *Proceedings of AAAI*, pages 691–696.

Kipper Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis.

Korhonen, A., Krymolowski, Y., and Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC*, volume 6.

Levin, B. (1993). *English verb classes and alternation, A preliminary investigation*.

Lippincott, T., Rimell, L., Verspoor, K., and Korhonen, A. (2013). Approaches to verb subcategorization for biomedicine. *Journal of biomedical informatics*, 46(2):212–227.

Majewska, O., Vulić, I., McCarthy, D., Huang, Y., Murakami, A., Laippala, V., and Korhonen, A. (2017). Investigating the cross-lingual translatability of verbnet-style classification. *Language Resources and Evaluation*.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Palmer, M., Kingsbury, P., and Gildea, D. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Peterson, D. W., Boyd-Graber, J., Palmer, M., and Kawhara, D. (2016). Leveraging verbnet to build corpus-specific verb clusters. *The SEM 2016 Organizing Committee*, page 102.

Pinker, S. (2013). *Learnability and cognition: The acquisition of argument structure*. MIT press.

Rios, M., Aziz, W., and Specia, L. (2011). Tine: A metric to assess mt adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 116–122. Association for Computational Linguistics.

Scarton, C., Sun, L., Schuler, K. K., Duran, M. S., Palmer,

M., and Korhonen, A. (2014). Verb clustering for Brazilian Portuguese. In *Proceedings of CICLing*, pages 25–39.

Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

Sun, L., Poibeau, T., Korhonen, A., and Messiant, C. (2010). Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of COLING*, pages 1056–1064.

Vulić, I., Mrkšić, N., and Korhonen, A. (2017). Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of EMNLP*.

Windisch Brown, S., Dligach, D., and Palmer, M. (2011). VerbNet class assignment as a WSD task. In *Proceedings of IWCS*, pages 85–94.

Appendix: Classification Guidelines

The annotators were presented with the following classification guidelines, along with the list of 267 verbs in their native language, at the start of the task:

Here is a long list of verbs, with one verb in each line.
Please put them together in groups where you feel they are used to express similar or related meanings. For example you may feel 'throw, kick, punch' are related, or 'speak, talk, and write'. These groups can be broader (more members) or narrower (fewer members) but any group must have at least 3-5 members. Aim for cohesive small groups if possible and you can add a 'relationship link' from each group to any other groups if you feel there are relationships between the two groups. The relationship could be similar-to (bidirectional) or broader-than (unidirectional). Any verbs you cannot find a good place for, please put in a 'Miscellaneous' group. There is no problem with putting a verb in more than one class if it fits all, for example because a verb may have several different meanings.
We suggest using Microsoft Excel or a related spreadsheet program (e.g. Google Sheets) to constantly have an overview of current groups. The expected output is: (i) groups of verbs according to your own criteria (see above), (ii) relationship links between groups as also discussed above. To facilitate the linking, you can provide simple labels for each group, e.g., Group 1, Group 2.
There is not necessarily a fully correct solution to this task and a perfect grouping. It is perfectly reasonable to use your intuition or gut feeling as a native speaker while working on this task.